Les sources de l'historien à l'heure d'Internet

Philippe RYGIEL*

In Hypothèses, 2003, 341-354

> « L'enseignement de l'écriture, ô roi, dit Theuth, accroîtra la science et la mémoire des Égyptiens ; car j'ai trouvé là le remède de l'oubli et de l'ignorance. » Le roi répondit [...] « Toi, père de l'écriture, tu lui attribues bénévolement une efficacité contraire à celle dont elle est capable ; car elle produira l'oubli dans les âmes en leur faisant négliger la mémoire : confiants dans l'écriture, c'est du dehors, par des caractères étrangers et non plus du dedans, du fond d'eux-mêmes, qu'ils chercheront à susciter leurs souvenirs ; tu as trouvé le moyen, non pas de retenir, mais de renouveler le souvenir, et ce que tu vas procurer à tes disciples c'est la présomption qu'ils ont la science, non la science elle-même ; car, quand ils auront beaucoup lu sans apprendre, ils se croiront très savants. »
>
> Platon, *Phèdre*

Les historiens utilisent peu Internet, du moins leurs notes mentionnent-elles rarement des publications électroniques ou des ressources documentaires fournies par le réseau. Plus rares encore sont ceux qui construisent leur appareil de sources en utilisant les possibilités du réseau. Une étude américaine, examinant les références et renvois mobilisés par les articles parus dans quelques grandes revues américaines établit qu'en 2000 moins de 3 % des articles parus renvoyaient à des documents numérisés présents sur le réseau[1].

Il y a donc quelque provocation à affirmer que cet outil nouveau peut contribuer à transformer les conditions de travail des historiens et partant leur outillage intellectuel et leurs productions. Il nous faut pour le comprendre examiner les ressources auxquelles le réseau permet d'accéder,

---

* Université Paris I Panthéon-Sorbonne ; directeur de publication des *Actes de l'histoire de l'immigration* (http://barthes.ens.fr)/clio), responsable du site Internet du Mouvement social.

1. S. GRAHAM, « Historians and Electronic Resources : A Second Citation Analysis », *Journal of the Association for History and Computing*, vol. IV, n° 2, août 2001, http://mcel.pacificu.edu/JAHC/JAHCiv2/ARTICLES/ArticlesIV2.HTML. Les url indiquées dans ce texte sont actives à la date de sa rédaction, soit en septembre 2003.

ceux aussi qui seront distribués par ce biais dans un futur proche, et examiner les conditions de leur appropriation par l'historien, soit les modes de recherche d'information adaptés à cet univers de ressources, et tant les difficultés posées par le document numérique que les possibilités qu'ouvre son utilisation.

Les ressources du réseau

Il nous faut, au préalable, préciser deux points de vocabulaire, qui ne sont pas sans conséquences sur le contenu de notre exposé. Internet est un ensemble de réseaux informatiques interconnectés par le biais d'un protocole TCP/IP qui permet que transitent entre des machines très hétérogènes des données numériques, soit de l'information au sens que les physiciens donnent à ce terme. Traiter des sources et d'Internet c'est donc d'abord évoquer, du point de vue de l'historien, les caractéristiques des données numériques. D'autre part, du sens que nous donnons au mot source découle qu'aucun historien n'a jamais trouvé de sources sur le réseau, ni n'en trouvera. Cette attente, fréquemment exprimée par des étudiants, et forcément déçue, repose sur l'oubli d'une dimension fondamentale du travail de l'historien. La collection de sources mobilisée par une étude historique est le produit d'un travail – nécessitant de solides compétences et une bonne connaissance de son matériau – de sélection, de validation et de classement de documents, et nous prenons ce dernier terme dans un sens très large qui englobe toute trace matérielle d'une activité humaine dont l'historien entreprend l'étude. La source donc doit être construite et non trouvée, et ce travail exige un temps et une culture professionnelle que jamais un réseau ne saurait abolir, même s'il transforme certaines des propriétés de cette durée et exige certaines compétences spécifiques. Notons de plus qu'une telle définition exclut de fait les bases de données documentaires complexes dont les notices incorporent des informations puisées à plusieurs sources primaires, tels les répertoires prosopographiques qui fleurissent aujourd'hui sur Internet[2], que nous assimilons, par convention, à des éléments de bibliographie et non aux matériaux bruts du travail de l'historien, même s'ils peuvent évidemment être utilisés afin de constituer une source.

Listes, cartes et plans

Internet offre aujourd'hui à l'historien plusieurs types de documents utilisables lors de ce travail. Nous y trouvons de nombreuses listes permettant de repérer les coordonnées géographiques et institutionnelles d'objets utiles. Les musées, les bibliothèques, les centres d'archives, voire les maisons d'édition ou les revues sont de plus en plus nombreux à offrir des répertoires ou des catalogues. Ceux-ci tendent à devenir des répertoires

2. Le projet *Bright sparcs*, un dictionnaire des scientifiques australiens passés et présents en constitue un exemple abouti, http://www.asap.unimelb.edu.au/bsparcs/bsparcshome.htm.

dynamiques. Le client, en l'occurrence l'historien, peut présenter, à l'aide d'un formulaire, une requête et recevoir en retour une réponse, qui prend généralement la forme d'une liste de fiches répondant aux critères formulés, retrouvées dans une base de données à laquelle l'historien n'a pas accès. Cette organisation, indispensable dès lors que la quantité de données détenue devient importante, permet de formuler des requêtes extrêmement précises et complexes, puisque l'on peut poser à l'aide d'opérateurs booléens des conditions portant sur plusieurs rubriques de la base de données. Il m'est ainsi possible, par le biais du catalogue informatisé des collections d'archéologie, d'art contemporain, d'arts décoratifs, d'histoire, de beaux-arts et d'ethnologie appartenant aux musées de France de repérer les représentations de Louis XVI visibles dans l'un des musées nationaux de Paris et d'obtenir pour chacune un descriptif sommaire, voire une image, et ses coordonnées[3]. Je puis ainsi, cherchant la réponse à une question ponctuelle, retrouver rapidement la trace d'un document utile, ce à quoi j'aurais peut-être renoncé s'il m'avait fallu localiser puis consulter un ou plusieurs volumineux répertoires-papier.

Revers de la médaille cependant, il faut, pour accéder à l'information désirée, emprunter des chemins tracés par d'autres, qui ont choisi des rubriques, des modes d'indexation et construit des thesaurus en anticipant des besoins qui ne sont pas forcément les nôtres. Je ne puis ainsi, utilisant la même base Joconde, accéder de façon immédiate à une liste des estampes en taille douce publiées à Paris à la fin du XVIII^e^ siècle, alors même que le procédé de fabrication et le lieu de parution des documents archivés figure dans la fiche qui leur est consacrée ; mais l'auteur de la base de données n'a pas jugé utile de permettre une interrogation sur ces rubriques, ce qui contraint l'utilisateur à demander la liste de toutes les estampes de la période qu'il étudie, puis à consulter une à une chaque fiche détaillée afin de retenir celles qui sont conformes à sa recherche. La procédure n'est sans doute pas beaucoup plus longue que s'il nous fallait utiliser un document-papier pour faire le même travail, mais elle n'apporte pas de gain de temps significatif. De façon plus générale, il nous faut maîtriser, si nous voulons obtenir une réponse, le lexique et la syntaxe de la base de données, ce qu'il est généralement possible de faire à partir des indications fournies par le site, mais qui peut nécessiter un certain temps. Il nous faut ainsi savoir, reprenant le même exemple, que la rubrique désignant le type de document recherché (estampe, tableau, etc.) porte le nom de « domaine ». Cette connaissance de l'organisation de l'information est sans doute moins nécessaire, ou plus immédiate parce que correspondant aux habitudes héritées de notre formation, pour qui utilise un répertoire-papier, que nous avons toujours le loisir de parcourir en sa totalité afin de nous l'approprier.

La numérisation de l'archive

3. http://www.culture.fr/documentation/ccmf/pres.htm.

De plus en plus souvent, nous parvenons à accéder non pas seulement à un répertoire d'objets, mais à une collection d'images de certains de ceux-ci. Celles-ci sont le fruit d'une multitude d'initiatives émanant d'acteurs très divers et sont souvent sans lien ni coordination entre elles. Des chercheurs, isolés ou membre d'un collectif, numérisent des collections de documents, qu'ils les détiennent ou non[4]. Des institutions muséales, des bibliothèques[5], des centres d'archives[6], des collectivités territoriales, des sociétés privées, voire de simples particuliers contribuent ainsi à l'enrichissement du réseau, et permettent un accès, généralement depuis le *world wide web*, aux données ainsi constituées. Ce mouvement prend aujourd'hui de l'ampleur, d'autant qu'il est dans de nombreux pays, dont la France, soutenu par une volonté politique, au nom d'une nécessaire numérisation du patrimoine savant et culturel de la nation, ou du besoin de permettre l'accès à des documents fragiles. Les projets de numérisation du patrimoine en cours en région Nord-Pas-de-Calais bénéficient ainsi d'importants financements publics. Près de 8 millions de francs ont été investis entre 1997 et 2000, provenant pour moitié de fonds européens, pour un quart de l'État et pour un quart des collectivités territoriales[7].

Il est donc probable que les collections de ce type soient, dans les années à venir, de plus en plus nombreuses et de plus en plus riches. Il n'est pas certain cependant que l'effort se poursuive indéfiniment, ni que l'accès à celles-ci demeure gratuit. De telles opérations ont en effet un coût, et celui-ci est important, ce que l'exemple d'une entreprise de qualité, mais limitée dans son ambition, illustrera. Les historiens du *Reading Area community college* se sont engagés en 1997 dans un projet dont la finalité était la numérisation et la mise en ligne d'un fonds documentaire relatif à la construction d'un canal ouvert en 1825. Il a fallu pour rendre accessible ce fonds, riche de 1 100 documents comprenant des plans, des diagrammes, des croquis, embaucher trois personnes qui ont, durant plusieurs mois, sous la conduite d'un historien consacrant une bonne partie de son temps à l'avancement du projet, numérisé les documents, construit la base de données les répertoriant

4. L'Institut de Recherche et d'Histoire des Textes (IRHT) poursuit depuis plus de vingt ans le recensement des manuscrits enluminés des bibliothèques de France et permet l'accès à une banque d'images couplée à une base de données textuelle offrant la possibilité de rechercher et de visualiser une partie des enluminures recensées.

5. La bibliothèque numérique de la BnF, Gallica, qui propose un accès à 70 000 ouvrages numérisés, à plus de 80 000 images et à plusieurs dizaines d'heures d'enregistrements sonores, demeure, dans le monde francophone, l'exemple le plus spectaculaire de telles entreprises : http://gallica.bnf.fr/.

6. Les archives départementales des Yvelines offrent ainsi l'accès à une collection de documents numérisés qui comprend surtout des cartes et plans anciens.

7. « Évolution du plan de numérisation des fonds d'État, bilan pratique 1996-1999, nouvelles orientations pour 2000-2001 », dans *La documentation numérique au ministère de la Culture et de la Communication*, 19/1/1999, http://www.culture.fr/culture/mrt/numerisation/fr/seminaire_191199/table_ronde1.htm.

et le site web permettant l'accès à celle-ci[8]. Le succès de l'entreprise suppose de plus que cet établissement dispose d'un serveur dont la maintenance soit efficacement assurée. Sa pérennité dépend, quant à elle, de l'entretien du site et en particulier du fait que soit assurée à intervalles réguliers une migration des données vers de nouveaux formats, ceux que traiteront les ordinateurs futurs, puisque que la durée de vie des formats informatiques est généralement assez brève, ce que savent tous ceux qui disposent d'archives numérisées vieilles de plus de dix ans périodiquement confrontés à la quasi impossibilité d'ouvrir d'anciens fichiers sur des machines récentes. Toute entreprise de numérisation implique donc un débours initial important, plus du fait d'ailleurs de la nécessité d'un travail humain que des dépenses liées à l'achat de matériel et génère des coûts de maintenance. L'une des questions qui se pose alors – et les enjeux en sont importants – est de savoir qui paiera et dans quelles proportions. Un désengagement des établissements publics ou un recours accru aux redevances d'utilisation risquent d'orienter l'effort de numérisation vers des collections pour lesquelles existe une demande solvable, ou une forte demande institutionnelle, voire dont la sélection répond à des nécessités techniques de conservation, qui ne sont pas nécessairement celles qui intéressent le plus l'historien.

Malgré ces incertitudes et ces limites, de telles entreprises sont utiles aux historiens ; il nous faut cependant là encore revenir à la matérialité des opérations de construction de ces documents pour mieux en apprécier la portée et les limites. Numériser un document consiste non à en prendre un cliché ou une empreinte, même si cela peut constituer la première étape de la numérisation, mais à produire, à partir de celui-ci, un document numérique, constitué d'une séquence de nombres binaires, générée par un opérateur humain qui a le choix entre plusieurs outils et plusieurs stratégies de numérisation, ainsi qu'entre plusieurs possibilités quant aux modes de classement et d'accès possibles aux documents produits. Lorsque le document est disponible par le biais du réseau, cette suite est adressée à l'ordinateur de l'historien et d'autres séquences binaires, comprenant les instructions des programmes dont il dispose sur sa machine, sont mobilisées afin de traiter le fichier reçu et d'afficher sur la surface de son écran une image lisible par lui. Nous n'assistons donc pas, consultant un site, à l'affichage d'un document d'archive, mais nous obtenons une représentation de celui-ci, qui résulte d'un processus d'abstraction dont les caractéristiques fixent les usages possibles de l'objet, qui sont en nombre infini, mais définis par les opérations le produisant. La nécessité de préciser ce point ne vient point d'un souci pédant d'user d'un vocabulaire exact. C'est en effet la connaissance de la genèse de l'objet que nous percevons qui nous permet de

8. J. M. JR LAWLOR, « The Schuylkill Navigation System Project At Reading Area Community College : Preservation and Dissemination of an Important Collection of Transportation Documents », *Journal of the Association for History and Computing*, V/1 (mai 2002) ; http://mcel.pacificu.edu/JAHC/JAHCV1/ARTICLES/lawlor/lawlor.html.

mieux comprendre ce que nous pouvons – et ne pouvons pas – en faire. Rien de choquant d'ailleurs dans cette affirmation selon laquelle des propriétés matérielles des objets manipulés par l'historien et de celles des contextes d'usage dépendent les opérations qu'il peut effectuer, et donc, *in fine*, certaines des caractéristiques de l'objet de connaissance qu'il produit. Il suffit pour s'en convaincre de se référer aux travaux des historiens du livre[9].

En l'occurrence cette attention aux propriétés de l'objet manipulé permet d'abord de comprendre que la consultation d'un site n'est pas du même ordre que l'examen d'un document d'archive et ne peut, dans certains cas, se substituer à elle. La représentation obtenue, parce qu'elle est abstraction, ne conserve pas certaines des propriétés sensibles du document numérisé : ni sa texture, ni son odeur, ni sa structure moléculaire, pour ne prendre que quelques exemples, ne nous sont accessibles, ce qui constitue, dans le cadre de certaines recherches, une perte sensible – pensons, par exemple, à un historien qui étudierait les techniques de production des manuscrits anciens.

D'autre part, la page-écran à laquelle nous accédons incorpore des choix faits par les opérateurs chargés de la numérisation et de la mise à disposition du document, choix qui ne sont techniques qu'en apparence puisqu'ils enchâssent des représentations des usages légitimes ou souhaités des documents disponibles, ainsi que des compétences possédées par les utilisateurs et déterminent les usages possibles. Prenons pour illustrer ceci l'exemple simple d'une circulaire d'une page adressée par un préfet aux maires de son département. Celle-ci peut-être numérisée en mode image, nous aurons alors accès à l'équivalent d'un cliché du document, qui préservera sa disposition matérielle, mais ne permettra pas de manipuler le texte de la circulaire. Il sera alors impossible, par exemple, de la citer sans recopier le passage pertinent, ou de chercher un mot dans le texte sans lire celui-ci entièrement. Le mode texte à l'inverse le permettra, mais nous ne pourrons, nous posant des questions sur la date du texte ou son authenticité, ni examiner sa disposition matérielle ni son aspect. Si nous choisissons le mode texte, il nous faut ensuite savoir si nous le présenterons en utilisant soit l'Html, qui est le format le plus courant sur le web, mais ne permet pas de fixer l'aspect de la page-écran qu'observera l'internaute (les caractères, les couleurs, la taille des colonnes varieront d'une machine à l'autre), soit l'Xml, qui permet d'incorporer au document des métadonnées décrivant avec précision la structure de celui-ci, ce qui en facilite l'incorporation dans une base de données et l'indexation, soit le PDF, qui est un format propriétaire qui contraint le visiteur à quelques manipulations de plus, ce dont nous pouvons, en fonction de ce que nous savons, ou devinons, du public futur, le croire ou non capable, mais qui a l'avantage de

9. Par exemple *Histoire de la lecture dans le monde occidental*, G. GUGLIELMO, R. CHARTIER dir., Paris, 1997.

garantir une impression du document à l'identique, quelle que soit la machine utilisée, et donc offre la possibilité de reproduire assez fidèlement la disposition du texte original.

En somme, et nous avons là l'une des explications du relatif dédain dans lequel les historiens semblent tenir les ressources numériques, nous sommes en permanence confrontés à un écrit-écran dont les caractéristiques ont été fixées par d'autres, qui sont rarement des historiens et prennent rarement en compte les besoins des historiens, d'ailleurs très divers d'une recherche à l'autre. En d'autres termes, il faut, le plus souvent, ou retravailler un matériau qui n'a été produit ni par ni pour les historiens, ou accepter que celui-ci ne réponde qu'à certaines de nos attentes et ne dispense pas toujours de retourner à la source de l'objet examiné.

L'archive numérique

Ces conclusions sont également valables si nous examinons les données n'existant que sous forme numérique, qu'il s'agisse des contenus du *web*, des archives des listes de diffusion, des *newsgroups* ou des données numériques produites par les entreprises et les administrations qui, de plus en plus, n'ont pas d'équivalent papier. Ce monceau de données constitue un ensemble extrêmement riche et volumineux de dispositifs discursifs qui constituent un matériau de choix tant pour l'historien du très contemporain que pour l'historien de demain, à condition cependant qu'émergent des dispositifs d'archivages permettant de pallier leur volatilité. Ces données en effet ne sont disponibles, et ne peuvent éventuellement être accessibles par le réseau que tant qu'elles sont physiquement inscrites dans une mémoire de masse à laquelle il est possible d'accéder et codées en un format lisible. Or cet ensemble de conditions est fort difficile à réunir. Nous avons déjà évoqué la faible durée de vie des formats informatiques, ajoutons que nombre de supports informatiques ne sont guère plus pérennes, les disques magnétiques s'effaçant plus vite que l'encre des livres. De plus, dans le cas des sites *web*, les opérateurs, qu'ils s'agissent des auteurs des sites ou de ceux qui en assurent la sauvegarde ou la mise à disposition, peuvent rapidement disparaître ou abandonner un site. Enfin, les documents mis à disposition du public par le biais du réseau – et en cela la métaphore de la page se révèle très inexacte – ne sont pas figés, mais au contraire fréquemment modifiés, corrigés et mis à jour sans qu'il ne reste la plupart du temps de trace des états antérieurs.

La tâche est donc rude et coûteuse d'autant que, pour le seul cas du *web*, la croissance du volume des données portées sur le *web* est exponentielle – il naîtrait aujourd'hui de 4 à 8 millions de pages par jour tandis que

plusieurs millions disparaissent ou se modifient chaque seconde – alors que la technologie du stockage semble actuellement rencontrer un palier[10].

Ces difficultés sont anciennes et connues de la plupart des acteurs concernés dont beaucoup ont, dans un certain désordre, entrepris d'archiver tout ou partie des données présentes sur le réseau ou des données numériques vitales. Plusieurs acteurs privés ont entrepris d'archiver tout ou partie du réseau pour leur propre compte. L'*Internet Archive* de Scott Kirkpatrick génère de petits robots informatiques qui prennent des instantanés de la toile, ce qui lui permet d'affirmer disposer de plus de onze milliards de pages *web* archivées et parvenir à en archiver plus de 120 millions par jour[11]. L'autoarchivage quelque peu décentré et anarchique du réseau – sans compter la possibilité que seuls des organismes privés puissent posséder la mémoire de parties de celui-ci – ont conduit plusieurs gouvernements à mettre en place très récemment des dispositifs d'archivage publics des données numériques et des informations disponibles sur Internet. Au Canada, le discours du trône du 30 septembre 2002 a créé une institution du savoir qui a pour fonction d'offrir aux Canadiens un « accès facile et intégré à leur patrimoine documentaire et au savoir sur la société canadienne », ce qui comprend entre autres missions le prélèvement périodique d'échantillons de site canadiens et de sites traitant du Canada[12]. En France, une loi datée du 13 juin 2001 prévoit elle aussi l'archivage automatique de sites *web*, mission confiée conjointement à la BnF et à l'INA. La mise en place de tels dispositifs est cependant loin de résoudre toutes les difficultés, ne serait-ce que parce que les sites dynamiques – qui se généralisent – peuvent difficilement, en l'état actuel des choses, être automatiquement aspirés, ou parce que le cadre juridique de la mise à disposition de ces données n'apparaît pas définitivement fixé[13].

Les données numériques non textuelles

Nous avons jusqu'ici évoqué des données susceptibles de donner naissance à une visualisation pouvant faire sens pour tout utilisateur, pour peu qu'il sache se servir d'un navigateur et maîtrise les codes linguistiques et sémiotiques en usage dans le monde contemporain. L'activité du réseau génère de nombreux autres documents, eux aussi virtuellement source. Nous n'en donnons ici qu'un exemple, les lecteurs intéressés pouvant se

10. V. FARISSON, « La mémoire de l'Internet », intervention au *Séminaire Internet de Sciences-Po* (Paul Mathias), 2000-2001, http://barthes.ens.fr/scpo/Presentations00-01/Farison_MemoireNet.html.

11. Ce site permet aussi l'accès à des collections d'images, de livres, de films et de documents sonores numérisés. Destiné à servir la recherche et les historiens, il permet la consultation gratuite de ses fonds, http://www.archive.org/index.php .

12. http://www.nlc-bnc.ca/10/11/a11-300-f.html.

13. Pour les aspects juridiques des entreprises de numérisation et de mise à disposition de produits numérisés sur le réseau voir *La numérisation pour l'enseignement et la recherche. Aspects juridiques*, I. de LAMBERTIE dir., Paris, 2002.

reporter aux travaux d'Éric Guichard[14]. Les utilisateurs des machines abritant des données accessibles depuis le réseau ont la possibilité d'enregistrer de multiples informations concernant la provenance et les pratiques de ceux qui les utilisent, voire leur identité et ce, très souvent à l'insu de ceux-ci. Les fichiers logs ainsi constitués, qui ne sont pas des fichiers en langue naturelle susceptibles d'une lecture immédiate, permettent de multiples traitements, qui ont généralement une finalité pratique : l'optimisation des sites hébergés (évaluation de la bande passante nécessaire, refonte de l'architecture afin de mieux satisfaire les besoins des visiteurs), ou commerciale (il est possible de se constituer ainsi à peu de frais un fichier d'adresses électroniques, associant à chacune de celles-ci un profil de consommateur que l'on revendra aux spécialistes du marketing direct ou que l'on utilisera pour promouvoir auprès d'eux ses propres produits). Ces données peuvent cependant être mobilisées par le chercheur à d'autres fins, devenant des indicateurs de l'activité des pôles universitaires français[15], permettant d'évaluer la provenance géographique des visiteurs d'un site, voire, lorsque nous pouvons étudier les requêtes formulées par les internautes utilisant un grand moteur de recherches francophone, nous donnant de précieuses indications sur les modes d'appropriation et de lecture du *web* des internautes français et de langue française[16]. Là encore, les historiens poussant leurs investigations jusqu'au très contemporain peuvent trouver du grain à moudre, à condition toutefois d'être capables de localiser ces sources et d'extraire et traiter l'information issue de tels fichiers.

Les méandres du net

De façon plus générale l'usage, entendons l'usage professionnel, d'Internet n'est jamais ni immédiat ni intuitif ni transparent, il suppose l'acquisition de compétences (nous en avons déjà évoqué quelques-unes) que n'acquièrent pas traditionnellement les historiens[17] et dont l'apprentissage suppose du temps – et cela même lorsque nous envisageons des tâches en apparence simples, telles que chercher une information pertinente et utilisable sur Internet. L'un des commentaires les plus fréquents faits par les étudiants, ou les historiens, utilisant Internet est en effet que le réseau regorge sans doute d'informations mais qu'elles sont fort difficiles à trouver et que leur fiabilité est sujette à caution. La structure du réseau l'explique en partie : aucune autorité centrale ne contrôlant ni ne

14. É. GUICHARD, *L'Internet : mesures des appropriations d'une technique intellectuelle,* Thèse de doctorat de l'École des Hautes Études en Sciences Sociales (option sciences de l'information et de la communication), 2002, http://barthes.ens.fr/atelier/theseEG/theseEG.html.
15. ID., « Cartographie animée du réseau Renater » *Atelier Internet*, 2001, http://barthes.ens.fr/atelier/geo/Renater01/.
16. ID., *L'Internet : mesure des appropriations, op. cit.*
17. J.-P. GENET, « La formation informatique des historiens en France : une urgence », *Mémoire vive*, 9 (1994).

classant les documents accessibles depuis celui-ci, le réseau ressemble de ce fait physiquement à une gigantesque collection de greniers, parfois rarement nettoyés, en lesquels de multiples individus déposent des documents de toute nature – à ceci près que nous avons immédiatement accès à une bonne partie de ces lieux de dépôts, qu'il est souvent possible d'écrire au propriétaire des lieux et que les documents qu'ils recèlent sont constitués de formes (en l'occurrence des séries numériques) que les utilisateurs des machines du réseau peuvent identifier et manipuler, ce qui permet l'indexage et le catalogage d'une partie des ressources disponibles. Aucune instance centrale n'assure, même à l'échelon local, ces tâches : les bibliothécaires et les documentalistes du réseau sont fort nombreux, même s'ils sont loin de suffire à la tâche, et extraordinairement divers ; des chercheurs, de simples particuliers, des acteurs institutionnels, des compagnies privées tissent les cartes entrecroisées de multiples territoires virtuels et nous sont des auxiliaires indispensables. Leur travail aboutit à la production de deux types d'objets favorisant la navigation, des portails et des moteurs de recherche, dont il faut généralement combiner les indications dans le cadre d'une recherche.

Les sites-portails sont construits par des spécialistes d'un domaine, des institutions, des sociétés commerciales ou de simples passionnés et recensent des ressources utiles aux spécialistes et aux étudiants, généralement classées, dans le meilleur des cas présentées et évaluées. Ils constituent de bons points de départs pour une recherche documentaire. Je ne puis, dans le cadre de cet article, citer tous les portails francophones utiles aux historiens en quête de sources, et je renvoie donc au recensement des sites portails d'histoire effectué par Christine Ducourtieux pour l'École doctorale de Paris I[18], me contentant de signaler que les médiévistes disposent avec Ménestrel d'un très bel outil de ce genre[19] et que tous les sites des archives départementales sont recensés par un portail destiné aux généalogistes[20].

Il est souvent nécessaire de croiser entre elles les informations fournies par ces sites et de les compléter par celles obtenues par le moyen d'un moteur de recherche, de la même façon qu'un historien approchant un sujet est contraint de compulser plusieurs répertoires de sources et de nombreux ouvrages. Contrairement aux portails, la plupart des moteurs de recherche ne sont pas conçus par et pour des historiens, mais sont créés par des sociétés commerciales et destinés à un large public. Ils reposent tous sur des principes assez similaires. Un moteur indexe automatiquement les formes verbales, en faisant abstraction de leur sens, que ses robots rencontrent sur les pages. Ce fonctionnement permet aux moteurs de

18. http://edoc-histoire.univ-paris1.fr/formations.htm.
19. La page Ménestrel consacrée aux sites *web* des centres d'archives se trouve à http://www.ccr.jussieu.fr/urfist/menestrel/paleo/paleo-04archiv.htm.
20. http://www.memodoc.com/liste_archives_departementales.html.

recherche de renvoyer à l'internaute un grand nombre de réponses sur tous types de requêtes, à partir du contenu de la totalité des pages stockées dans leur index. Le classement des résultats se fait selon un algorithme de pertinence vis-à-vis de la requête formulée, chaque moteur ayant son propre algorithme. Il s'en déduit, et l'étude des fichiers logs des moteurs permet de le confirmer[21], que la quasi-totalité des requêtes obtiennent des réponses non pertinentes et inutilisables car, constituées de peu de mots et de mots peu discriminants, elles génèrent des listes interminables où figurent beaucoup de pages inutiles au chercheur et, celui-ci consultant rarement plus d'une ou deux des pages concernées, il est fréquent que le moteur retourne une information pertinente dont l'internaute ne prenne pas connaissance. Il faut donc, pour pouvoir utiliser avec efficacité ces outils, développer des stratégies adaptées à son objet, et en particulier formuler des requêtes longues, si possibles composées de plusieurs termes les plus discriminants possibles, voire lancer plusieurs requêtes, légèrement différentes les unes des autres, adressées à plusieurs moteurs, et avant d'explorer la liste des réponses examiner les indications de provenance et de contenu que le moteur retourne. La requête « états généraux, Necker, Louis XVI, trône, discours » me permet ainsi, grâce à un enseignant de l'université de Hessen qui a porté en ligne l'essentiel de son cours et des documents utilisés dans le cadre de celui-ci, de récupérer de larges extraits du discours d'ouverture des États généraux, la référence à ce document arrivant en sixième position sur la liste des réponses fournies par un moteur de recherche. J'ajoute que ce document est le premier de ceux de la liste que j'ai consultés, tant l'adresse des autres sites que les autres indications fournies par le moteur me faisant douter de pouvoir y trouver ce que je cherchais.

Validation de l'information

J'ajoute qu'ayant spontanément confiance en la compétence et le sérieux d'un collègue, en celle aussi de l'administrateur-réseau d'une université allemande (que j'ai pu identifier grâce à son url), je suis assez raisonnablement certain du sérieux de ma source – du moins serais-je prêt à l'utiliser telle quelle dans le cadre d'un cours, alors que je vérifierais probablement la transcription et retournerais sans doute à l'original s'il s'agissait là d'une pièce importante du dossier documentaire d'une recherche en cours. Les choses cependant ne sont pas toujours aussi simples, particulièrement pour les contemporanéistes qui ne chercheraient pas la trace de sources anciennes mais voudraient constituer un dossier documentaire, concernant par exemple l'activité d'une organisation politique, incluant des documents Internet. En effet, la fiabilité des données accessibles depuis le réseau pose véritablement problème dans un certain nombre de cas. Deux raisons à cela. D'une part, nous ne disposons pas des repères physiques et institutionnels (sa présence dans un dépôt d'archives

21. É. GUICHARD, *L'Internet …*, *op. cit.*

par exemple) qui nous permettent d'habitude d'identifier le document que nous utilisons. D'autre part, du fait de sa nature même, l'information numérique est assez facile à copier et à modifier de façon indécelable par l'utilisateur – et s'introduire dans un serveur afin d'en modifier les données est à la portée de nombre d'informaticiens et de bricoleurs. Un groupe non identifié de libéraux furieux a ainsi attaqué il y a quelques mois le site de la revue *Le Mouvement Social*, les visiteurs accédant à une page blanche sur laquelle de grosses lettres noires indiquaient que ce site ne serait pas accessible tant que le mouvement social prendrait la France en otage. Un historien tant soit peu averti pouvait cependant déceler la modification et douter de ce que l'équipe de cette vénérable revue ait soudain décidé de se croiser afin de défendre la société libérale menacée par des hordes de grévistes. Boutade bien sûr, mais qui voudrait rappeler que les historiens sont parmi les mieux armés pour affronter cette incertitude et ne devraient point en être effrayés. Le document obtenu depuis Internet est, pour peu que l'on ait quelque idée de sa structure et de sa matérialité, susceptible comme tout document d'une critique tant interne qu'externe. Son adresse (ou url) donne ainsi des indications sur sa localisation physique et parfois sur l'inscription institutionnelle de l'auteur du document, l'affichage des sources de la page permet dans un certain nombre de cas de vérifier la date de modification affichée sur celle-ci ; reste enfin le loisir de croiser le document obtenu avec d'autres, qu'ils soient électroniques ou non. Ajoutons enfin que le réseau n'est pas parcouru sans relâche par des esprits diaboliques s'acharnant à falsifier les données intéressant les historiens.

Lire les sources numériques

Le lecteur pourra se demander à ce point, après l'évocation de tant de peines et de difficultés, si le coût de constitution d'un recueil de sources numériques est justifié. La réponse varie dans d'importantes proportions selon le type de sujet et de questions que traite l'historien. À la date d'aujourd'hui, les données numériques fournies par le réseau sont dans certains cas de modestes adjuvants, dans d'autres des sources indispensables. Il y a à cette conclusion deux prémisses. D'une part, l'ubiquité du réseau permet le rassemblement de données géographiquement dispersées à un coût et en un temps acceptables. D'autre part, les données numériques, même lorsqu'elle ont un équivalent physique, ne sont jamais des représentations appauvries d'un référent réel, mais un objet, tout aussi matériel mais d'une autre nature, qui rend possible des formes de lecture ou d'appropriation spécifiques : les fichiers numériques permettent en particulier la manipulation de gigantesques quantités de données et l'automatisation d'un certain nombre de tâches. De ce fait elles autorisent des questions qu'un historien ne pourrait poser sans elles.

Je me propose d'illustrer ce point en évoquant plusieurs recherches dont les résultats sont déjà publiés, chacune illustrant, sans que l'inventaire

soit exhaustif, un usage possible d'un appareil de sources constitué pour tout ou partie à l'aide du réseau en allant du plus simple, ou du plus familier, au plus complexe.

Le plus court chemin de l'historien à sa source

Une étude italienne consacrée aux modes d'autoreprésentation graphique adoptés par les partis socialistes et communistes européens, c'est-à-dire aux logos et symboles choisis par eux, me fournit mon premier exemple[22]. L'ouvrage se clôt par un dernier chapitre qui utilise pour sources les sites de nombreux partis socialistes et étudie les logos et les dispositifs graphiques de ceux-ci. Les outils d'analyse utilisés sont fort classiques et relèvent pour l'essentiel d'une sémiotique qui pourrait être mise en œuvre à partir de sources classiques, et l'est d'ailleurs dans la première partie de l'ouvrage. Ce qu'apporte ici Internet est la possibilité à moindre coût, et en un temps limité, de rassembler une documentation physiquement très dispersée et de prendre comme échelle d'observation le monde. Certes il ne serait pas impossible de se fixer comme objectif de récupérer une collection de tels objets par d'autres moyens. Il faudrait pour cela se procurer les adresses de toutes les organisations appartenant à cette famille, puis écrire à chacune d'elles en lui demandant de bien vouloir envoyer quelques échantillons de ses dispositifs graphiques. L'opération cependant a un coût, certes encore modeste en ce cas, demande quelques compétences linguistiques, implique un délai de réponse plus ou moins long, et il est fort probable que le taux de réponse ne soit pas de 100 %. En d'autres termes il est ici plus facile, plus rapide et moins coûteux de constituer l'appareil de sources à partir d'Internet, même s'il serait possible de le faire par d'autres moyens.

De nouveaux modes discursifs

Dans un autre registre, plusieurs historiens et spécialistes des sciences politiques s'intéressent aujourd'hui beaucoup aux formes discursives nouvelles nées avec la diffusion du réseau – sites *web*, mais aussi forum de discussion et listes de diffusion – parce qu'ils permettent d'avoir accès à des formes de communications internes au groupe ou aux réseaux sociaux, auxquelles l'historien a rarement accès : la conversation, l'échange téléphonique, voire souvent la feuille ronéotée, laissant peu de traces. Fabien Granjon a ainsi étudié les répertoires d'action de plusieurs réseaux et associations politiques appartenant à la mouvance anti-mondialiste, à l'aide entre autres sources d'archives numériques qu'il a lui-même constituées, stockant les messages passant par les forums et les listes de diffusion de plusieurs associations[23]. En ce cas encore les méthodes d'analyse utilisées

22. *Scrivere con la sinistra. Dalla carta intestata a Internet*, S. CARETTI, M. DEGL'INNOCENTI, G. SILEI dir., Rome, 2002.
23. F. GRANJON, « Les répertoires d'action du néo-militantisme », *Le Mouvement Social,* 200 (juillet-septembre 2002).

sont les mêmes que celles des historiens travaillant des documents-papier, mais la source ne peut cependant être constituée que grâce au réseau, et elle permet de décrire dans la durée les échanges informels au sein du groupe beaucoup plus finement que n'importe quelle autre source.

Analyse textuelle

Les données numérisées fournies par le réseau offrent de plus la possibilité de mettre en œuvre, sur de grands corpus, des modes d'exploration et d'utilisation de l'information automatisés. Ceux-ci sont nombreux mais nous ne mentionnerons que l'analyse textuelle, qui apparaît ici particulièrement adaptée, à la fois parce que le réseau offre de vastes corpus de texte déjà numérisés et parce que les outils et les méthodologies de ces formes d'analyse sont déjà rôdés et utilisés depuis longtemps tant par des historiens[24] que par des spécialistes de l'analyse littéraire[25], certains voyant même dans ces formes d'analyse le simple prolongement de très anciennes pratiques. Michel Bernard écrit ainsi que :

> « […] on a pratiqué une forme manuelle de la statistique textuelle depuis le XIX^e^ siècle […]. Il ne s'agit donc pas de l'irruption dans le champ littéraire d'un intrus exogène et imposé de l'extérieur mais de la rencontre assez naturelle entre les techniques de la recherche littéraire et des outils qui la facilitent […] »[26].

Et ce, même si ces outils permettent des questionnements nouveaux ou d'obtenir des éléments de réponse précis à des interrogations anciennes. Il est ainsi possible, écrit-il, de vérifier la date de première attestation de tous les mots d'un poème afin de se rendre compte de la part d'archaïsmes et de néologismes qu'il utilise : « Qui aurait fait une telle démarche en feuilletant les pages d'un dictionnaire étymologique ? »[27].

Cette même tension entre radicale nouveauté et continuité est au cœur de tout discours dédié au trinôme que constituent l'historien, ses sources et le réseau, ou l'ordinateur – ce qui revient au fond au même –, lequel discours hésite toujours entre la métaphore rassurante qui, rapportant au connu, permet d'apprivoiser l'objet nouveau – au risque de méconnaître les ruptures qu'il peut introduire – et l'annonce prophétique. Nous n'avons pas ici échappé à la règle puisque nous sommes tenté de conclure que l'intégration du réseau dans les stratégies de construction de sources ne change pas les fondements du métier d'historien, même si elle suppose l'apprentissage d'un certain nombre de compétences nouvelles. L'historien doit toujours constituer des cartes mentales des gisements de traces disponibles – ces traces fussent-elles traces de traces –, en évaluer la

24. D. I. GREENSTEIN, *A Historian's Guide to Computing*, Oxford, 1994.
25. M. BERNARD, *Introduction aux études littéraires assistées par ordinateur*, Paris, 1999.
26. *Ibid.*, p. 8.
27. *Ibid.*, p. 14.

pertinence, sans le secours parfois de l'archiviste ou du documentaliste – ce qui est l'occasion de mesurer ce qu'il leur doit –, à partir d'une critique qui suppose la compréhension de la genèse et du mode de conservation des documents recensés. Il doit enfin rendre compte de l'économie de ces vestiges en utilisant des moyens appropriés à leurs caractéristiques, soit savoir les lire et les interpréter. Dans la mesure cependant où certaines recherches peuvent désormais mobiliser des données numériques à une échelle sans précédents, il semble probable qu'un nombre croissant d'historiens soit appelé à mobiliser des technologies intellectuelles nouvelles, dont bon nombre ne sont sans doute pas encore nées, ou sont encore inconnues d'eux, ce qui ne saurait manquer d'avoir des effets sur les questions qu'ils pourront demain poser à leurs sources.